LOGIQUE ET ANALYSE

NOUVELLE SÉRIE

47e Année Mars – Juin – Septembre – Décembre 2004

185 – 186 – 187 – 188

TABLE DES MATIÈRES

Publication trimestrielle du Centre National Belge de Recherches de Logique

Vakgroep Wijsbegeerte, VUB, Pleinlaan 2, B-1050 Brussel, Belgium

Logique & Analyse **185–188** (2004), 85–121

A LOGIC FOR THEORIES IN FLUX

LÁSZLÓ PÓLOS AND MICHAEL T. HANNAN

Introduction

It is an elementary requirement for any symbolic logic that the inferential behavior of sentences, formulæ, depend on nothing but logical forms. Studying theory building in the social sciences led us to the conclusion that the inferential behavior of several kinds of natural-language sentences cannot be accounted for in terms of the logical form that usual first-order formalizations attribute to them. In this paper we present our latest attempt to provide adequate logical forms for all the relevant kinds of sentences. The language we present is a modification and extension of one presented earlier (Pólos and Hannan 2001, 2002). The modifications reflect our experience in using the logic in formalizing sociological theories, especially theories of organizational ecology (Pólos, Hannan, and Carroll 2002; Hannan, Pólos, and Carroll 2003a, 2003b, 2003c; Hannan, Carroll, and Pólos 2003a, 2003b). We found that "rule like" statements, generic sentences that express rules with exceptions, are broadly used, and that it is a mistake to interpret these sentences as universally quantified formulæ.

Encouragingly, perhaps, these experiences led us to conclusions rather similar to the ones Imre Lakatos arrived to in his seminal study of scientific research programs. (Lakatos, I. 1978.) Lakatos argued extensively that Popper's demarcation between science and pseudo-science paints an unrealistic picture of the actual practice of scientific research. The falsification of a scientific theory often does not persuade researchers to abandon it. This behavior appears to be odd, as Popper thought is was, only if one assumes that researchers think in terms of universal claims, for which falsification should have been a lethal blow. If, on the other hand, they take their claims as generic — rather than universal — it is sensible that they protect the core insights of a theory by building a protective belt around the core that can be used to explain away some of the challenges that, in the absence of this protective belt, would be interpreted as successful falsifications of the (core) theory. We found the task of building a logical model of this type of protective behavior challenging. If the conclusions that can be drawn from the core

of the theory might be withdrawn in the presence of some auxiliary assumptions from the protective belt, then they are presumptions, expectations only: they can not be used without further ado as premises of new arguments.

The logic of the argumentation appears to be nonmonotonic, and an adequate formalization should assign different logical forms to the premises and the conclusions derived from them. Further scrutiny revealed that premises originating from the core and from the protective belt of the theory exhibit still different inferential behavior, so in an adequate formal language their logical form has to be different too. To formalize the protective belt we need appropriate logical forms for the auxiliary assumptions. Such assumptions are not persistent parts of a theory, and the core of the theory does not claim whatever these assumptions express. The empirical validity of these assumptions is not scrutinized. We concluded that a third, intensional quantifier has to be added to the ones we introduced in our earlier attempts, the "normally" and "presumably" quantifiers.

In addition of the expansion of the logical constants we also extended the applicability of some of the constructions. In the present version, the nonmonotonic intensional quantifiers are no longer restricted to be in the outmost operator position of a formula. Embedded occurrences are allowed, too, to enable us to formulate definitions (universally quantified sentences) based on properties individuals *normally* have.

To accommodate these changes we redesigned the formal semantics. The revision is a conservative extension of our earlier efforts: if it is restricted to the language fragment of the earlier efforts (Pólos and Hannan 2001, 2002) this semantics validates the same inferences. However, it is capable of handling the considerable increase of complexity the new syntax required.

Logic and Theory Building

Empirical theories are rarely formalized in the strict sense. They are typically presented in a pseudo-formal language, i.e., an extension of a natural language with some field-specific mathematical formalisms. The lack of strict formalization allows for the nature of generality of these theories to remain hidden. Most general considerations in fact appear in the form of bare plural sentences such as those in a famous argument by Stinchcombe (1965): "Routines in young organizations are less well developed than in older organizations," or "Organizations with better developed routines have a lower hazard of mortality," or (the claimed liability-of-newness theorem) "Young organizations have a higher hazard of mortality than older organizations." These indeed general statements — but are they universal? Would these sentences be considered to be false if someone discovered a population of organizations in which young organizations have a lower mortality

rate than that the old organizations? No. These claims are general — but not universal.

Linguists were puzzled by the formal grammar of this type of sentence for a long time. Carlsson (1974), in his dissertation, first concluded that these sentences are intensional in nature. Subsequent research by Kratzer (1995) and Diesing (1995) concluded that there is a (hidden) generic quantifier in the logical structure of these sentences. Schubert and Pelletier (1988) showed that no context-independent extensional quantifier would assign the appropriate truth-conditions to these sentences. It seems to be natural to conclude that, if there is any quantifier, then it should be *intensional.* Carlsson (1995) argues that to account for the formal semantics of the generic sentences the semantic universe need to contain entities that can be best called as rules or regularities, since the truth conditions of generic sentences are normally not expressible in term of sentences about the individual instances.

The meaning of such generically quantified sentences can be approximated by saying that they express rules-with-exceptions. Such rules are not sufficient to derive certain truth of objects, but they still might be useful to shape what we *expect* of unknown individual instances (Veltman 1995). Such expectations might well be all that is available for arguments in the process of construction of a theory. The partiality of available information means that the truth-value assignments yield value gaps occasionally. Moreover, it also creates the possibility that we have only rules-with exceptions — not strict rules.

If the generality of empirical theories often appears in the form of generically quantified sentences, then, of course, the critical challenge of these theories might not be a simple attempt of falsification by counter-example. The possibility of exceptions, counter-examples, is already "priced in." Accidental, non-reproducible exceptions might be ignored as "mistaken measurements" or "historical accidents". For example, a study of organizational morality in a population of organizations that encounters a political revolution might easily yield a counter-example to the Stinchcombe claim of a liability of newness if the revolution suddenly wipes out all older organizations. Other kind of exceptions yields more serious theoretical implications. They might be explained away with the help of premises from the protective belt. Alternatively they might lead to the extension of the core theory. For example, one might find that populations to which only large or well-endowed organizations can enter show low mortality even among the young members. Repeated, systematic exceptions of this type sometimes lead to extensions of the core theory.

Theory Building

We treat theories as intensional objects, with extended histories. Such histories can usefully be seen as sets of theory stages. A theory stage is composed of a persistent part and an ephemeral one. The persistent part monotonically expands as the theory develops. It forms the backbone, the identity of the theory. The persistent part is still decomposable to an empirically testable, i.e. falsifiable component, and a not falsifiable component. This later part contains meta-considerations, definitions and, as will argue below, auxiliary assumptions. Following Lakatos (1978), we refer to the persistent, and empirically testable part of the theory as the *core*, but while the core for Lakatos was a constant set of causal stories in our rendering of the theory under development the core is not stabile, it grows occasionally. There is an implicit component, the desiderata, concerning implications (desired theorems and non-theorems). Desiderata linger around the persistent part of the theory, and the theorems actually derived (derivable) form a stage of the theory are regularly compared to it. If desirable theorems are not derivable or undesirable theorems turn out to be derivable, the theory under construction is challenged. On the other hand the (provisional) theorems need not belong to the persistent part of the theory.

The meta-considerations are those extra-theoretical issues that are treated as non-problematic. They include rules concerned with the structure of legitimate inferences: the logic of the theory. Even though these considerations are often implicit, the theory would be radically different if the logic is changed. Typically the meta-considerations also include various parts of mathematics, e.g., the calculus, set theory, and probability theory.

Definitions and causal stories, or explanatory principles, contain the substantive insights of the theory. These definitions and claims can be strict rules (universally quantified sentences) but may include generic sentences too. The latter typically take the form ϕs are normally ψs. Of course, if the persistent part only contains strict empirical rules (universal statements, that convey subject-specific insights), then any implications of such rules (theorems) also sit in the persistent part.

Auxiliary assumptions are claims that are introduced into an argument to link the causal stories and meta-considerations on the one side, and theorems on the other side, in cases where the argument would not go through without additional specification. For instance, the classical population genetics of R. A. Fisher and Sewell Wright required a specification of the assignment of mates in the sexual transmission of genes between generations. That is, an auxiliary assumption was needed. The chosen assumption was random mating. Because of their auxiliary nature, such assumptions are treated as subject to replacement by other such assumptions as needed. This suggests to us that auxiliary assumptions ought to be considered as rules-with-exceptions,

The ephemeral component is the set of predictions and explanations that depend on rules-with-exceptions. These implications are either individual sentences or generic sentences. They often take the form: ϕs are presumably ψs. If the theory building succeeds, then these predictions and explanations satisfy the desiderata. Even better, they might also yield some unexpected, potentially interesting, results.

Theory building is a process that moves from one theory stage to another. (We are going to define the notion of as theory stage formally below.) Moves are fueled by critical challenges to the earlier phase/stage of the theory. The already accepted explanatory principles remain intact, but new principles might be adopted. (This appears to be part of normal scientific activity, the conceptual framework remains intact but considerations are refined.) To figure out the appropriate response to a critical challenge requires that inferences be made, and these inferences are sound. But sound *according to what logic*?

Here we present two lines of arguments. First we consider some implications of Lakatos's story, second we analyze the consequences of using rules with exceptions as explanatory principles.

Lakatos (1978) investigated whether or not one can tell what is falsified by a (hypothetically) successful falsification attempt. He concluded that the Duhem-Quine thesis, at least in its weaker interpretation, is obviously correct, i.e., the falsification, the inconsistency between the predictions of theory (stage) and a fact, is better seen as the inconsistency of two theories, more precisely their respective theorems are inconsistent. To avoid inconsistencies a protective belt offers additional (auxiliary) assumptions, to explain the inconsistencies away. Now had the logic been monotonic, the auxiliary assumptions would be of no help. If the inconsistency is derivable form the more limited set of (core) assumptions it has to be derivable from any more extended set of assumptions too. In other words, if Lakatos's idea concerning the functioning of scientific research programs is correct, and we believe it is, then the logic of scientific argumentation is bound to be nonmonotonic.

If a nonmonotonic logic offers a successful method for dealing with the falsification problem, it can, perhaps, be used to solve another notoriously difficult issue in theory building: the unification problem. We claimed above that theories are intensional objects with extended histories. Since there is no reason to assume that theory development is always linear, these histories may occasionally branch. A given theory stage might sometimes be extended simultaneously in two directions. Such parallel developments yield several, potentially inconsistent theory stages, that we might call theory fragments. A classical first-order rendering of such fragments often yields obvious inconsistencies. A nonmonotonic rendering is a promising alternative. It may remove the inconsistencies, and offer some substantively interesting

insights as a bonus. Below will provide two instances of successful unifications that delivered both the removal of the inconsistency and the bonus.

If the core of the theory or the set of auxiliary assumptions contains rules-with-exceptions, then the logic in use cannot be the best-understood logic: classical first-order logic. This is because classical first-order logic is monotonic; and the inferences used in theory building follow a nonmonotonic pattern. When a theory gets expanded, new explanatory principles are adopted and some of the old predictions might vanish. It becomes possible that:

$$\Phi \models \phi \text{ but } \Phi \cup \Psi \not\models \phi.$$

In the last quarter century, several nonmonotonic logics were proposed mainly in computer-science and also in logic and formal linguistics. These nonmonotonic logics were typically fine-tuned to their assigned jobs. To find out if any of them is adequate for specifying the nonmonotonic reasoning in theory building we first have to consider carefully what an adequate logic would look like. Formalizing carefully the insights the argumentation in theory building is based on into a model theory offers an unbiased way to describe the specific reasoning patterns used. To provide an axiomatization of this logic might be a step to be done in the future, but we believe at the present it is an increasing number of applications should test first what does this type of reasoning delivers. We offer some of these applications in the present paper and a much larger body of formalizations will be available soon in Hannan, Pólos and Carroll (in preparation).

Some of the key insights that we want to formalize are the following.

- The available information in the process of theory building is typically partial. It is rarely possible to identify which of the possible worlds is the actual one. The best that can be done is to identify a (small) set of possible worlds that contains the actual world. An adequate formal semantics should allow for this type of partiality, which in turn means that some sentences should have the truth value "true" or "false" in a subset of possible worlds while the others have a third value, "unknown".
- Scientific rules are defaults, rules with exceptions.
- If arguments are formulated from rules with exceptions, then the specificities of the arguments matter.
- More-specific arguments override less-specific ones.
- Specificity differences are persistent, and new information cannot overrule established specificity orders. This is an important constraint because it indicates that extensional inclusion between the antecedents might not be the right way to characterize specificity differences of premises. As a theory develops, the partiality decreases,

and new theory stages might yield different relationships among extensions. What is needed to establish a clear specificity difference is the inclusion between extensions in all still-possible worlds. In other words, we need an intensional definition of specificity differences.

- Whether one argument is more specific than another depends either on factual information or on dependable empirical generalizations, which we call causal stories.
- Only the first causal story in an argument chain defines the specificity of the argument. (We offer motivation for this choice below when we define specificity orderings of regularity chains)
- Equally specific arguments pointing in opposite directions eliminate each other's predictions.
- Arguments that point in opposite directions but whose specificities are not comparable also eliminate each other's predictions.
- Theory building follows the principle of informational monotonicity, i.e. the core of the theory does not shrink, and it occasionally expands. Therefore, explanatory principles, causal stories, are not deleted, even when they are partially overruled. Definitions and meta-considerations are persistent as well.
- Certain operations in first-order logic, which rely detailed factual knowledge about the facts (such as modus tollens and contraposition), should not have a counterpart in the new logic.[1]

A Language for Theory Building

If a symbolic logic is to capture the argumentation in theory building adequately, then the logical form of its sentences should carry all the necessary information about their inferential behavior. This is the reason why the first task is to define a language that assigns different logical forms to sentences with different argumentative functions. For this reason we extend the language of first-order logic with three intensional quantifiers. We offer these three quantifiers with their respective formal semantics as candidates for the intensional quantifiers Kratzer and Diesing identified in the logical structure of generic sentences. We need to retain the language of first-order logic, because definitions, and meta-considerations are typically presented in terms of classical first-order formulæ. We extend this classical language by adding

[1] To prove that our approach delivers results in line with these insights goes beyond the scope of the present paper, but the interested reader can find some detailed proofs of this type in Pólos, Hannan and Kamps (1999). Even though both the language is more extended here and the semantics is a bit more intricate the relevant part of the argument works here in precisely the same way

a new intensional quantifier $\mathfrak{N}$, which stands for the expression "normally." Universal quantification typically operates on a formula whose main logical connective is the conditional (material implication). Similarly generic sentences prefixed with the intensional quantifier $\mathfrak{N}$ have the conditional as their main operator. We find it useful to assume that all of the generic sentences used in building a theory have such a conditional structure. In developing the formal language, we require that $\mathfrak{N}$ quantifies only conditional sentences.

We follow the lead of Veltman's (1996) in arguing that the sentences prefixed with the normally quantifiers do not tell us much about what the case is. Instead, they tell us about what the case is expected to be. Furthermore, these expectations are not defined in terms of mathematical expectations, being often used in situations where the mathematical expectations are not justified by the information available. The knowledge justifying rules-with-exceptions is not strong enough to tell about the individual instances. Such rules express regularities that shape our expectations. Expectations might turn out to be factually correct or not. Nonetheless, it would be misleading to express them just like facts.

One important difference between facts and rules-with-exceptions concerns reusability. Sentences expressed in the language of first-order logic can be re-used. That is, classical conclusions derived from first-order premises can be used as assumptions for further derivations. This is not the case for rules-with-exceptions. If an expectation is derived, at least in part, from such rules, then it is not re-usable directly. The lack of reusability comes from nonmonotonicity.

Provisional theorems (expectations derived from rules-with-exceptions) will be expressed by formulæ with the intensional quantifier $\mathfrak{P}$. The nonmonotonicity in theory building shows up in connection with these derived expectations. In particular, if a theory gets elaborated, then the derived expectations often change. What used to be expected in an earlier stage of the theory is no warranted as an expectation.

Formulæ with the normally quantifier $\mathfrak{N}$ provide a formal statement of insightful causal stories, the substantive assumptions that form the core of the theory.[2] Formulæ prefixed with the $\mathfrak{P}$ are conclusions that depend, in

[2] Insightful causal stories are typically not expressed in terms of probability distributions. Insights capture patterned behavior of individual instances. One might feel tempted to speculate that this has something to do with the fact that humans are masters of recognizing patterns while notoriously bad in making judgments about probabilities. Scientific reasoning is incomplete in the absence of insightful causal stories, because these causal stories are the ultimate source the ah-ha feeling, understanding.

part, on the causal stories.[3] These conclusions are ephemeral, and they do not belong to the core of theory.

To learn the implications of an argument built on rules-with-exceptions — what is presumably the case, it is often not sufficient to know only the causal stories. As Duhem (1906) pointed out, certain auxiliary assumptions are generally needed. These assumptions might take the form of some simplifying assumptions, descriptions of constraints, which make mathematical modeling possible, might carve out mathematical models, or provide the interface between the causal stories and the models. Sometimes these assumptions describe measurement instructions, operationalizations. *Auxiliary* assumptions sit halfway between the causal stories and the presumable consequences. They are persistent in an evolving theory because the desired theorems are not derivable in their absence. But the theory does not claim that they provide causal insights; in fact, they might not be true at all. For example population biologists who invoke random mating do not claim that this assumptions is an insightful description of the real world, on the contrary they might be more or less suspicious about the empirical validity of such and assumption.

In our previous efforts (Pólos and Hannan 2001, 2002) we focused on the $\mathfrak{N}$, and the $\mathfrak{P}$ quantifiers. We now think that it is essential to make clear the argumentative role of the auxiliary assumptions (they belong to the persistent part of a theory but not to the core, since they are not exposed to falsification attempts) by defining a separate logical form. To display their intermediate status we introduce a third intensional quantifier: $\mathfrak{A}$.

There is a further logical reason to claim a specific form for the auxiliary assumptions involving the nonmonotonicity we face in theory building. In classical FoL, we can deal with auxiliary assumptions by appending them to the set of theoretical premises. In other words, we can condition the argument on these auxiliary assumptions, due to the deduction theorem:

$$\Gamma \cup \{\phi\} \models \psi \Rightarrow \Gamma \models (\phi \rightarrow \psi).$$

However, this derivation does not hold generally for arguments that contain rules-with-exceptions. Therefore, auxiliary assumptions cannot be treated by conditionalization. We need some other way to treat auxiliary information. We designed the quantifier $\mathfrak{A}$ to play this role.

To summarize: causal stories, auxiliary assumptions, and presumptions (or provisional theorems) have a shared responsibility for nonmonotonicity.

[3] Veltman (1996) argued that expectations are tests that may succeed or fail in a given information state but that they do not contribute to the information content of the information state. Our present follows a somewhat similar intuition.

However, causal stories (as we construe them) are informationally monotonic, they remain intact as the theory expands, they keep contributing to the theory even when they are partially, or even completely overridden by more specific causal stories. These considerations set a methodological agenda. Only those generalizations should be added as causal stories to a theory for which the theorist is prepared to accept that they will remain assumptions of the theory. If there are doubts that they will be acceptable in the future stages of a theory, then they are not good enough for the status of empirical generalization.

Now we are going to define a language, starting with the language of classical FoL. Some of these definitions just recapitulate standard constructions, and we provide them only to avoid misunderstanding. To distinguish them from the definitions that introduce novel constructions we use the label "definition" only for the second ones. We add the three intensional quantifiers to express causal stories, auxiliary assumptions, and presumptions. We define two semantics for this language. First we give a possible-world semantics. Then we use this semantics to build models for theory stages, and we define the second semantics in terms of theory stages. Once this second semantics is given, we can define the logical consequence relation for this language, which completes the task of defining the nonmonotonic logic that we believe is suitable to formalize inferencing in theory construction.

Syntax

The language we define, which we call the language of theory building, is an extension of the language of classical FoL. We add three operators to the language, for Normally, Presumably, and Assumedly.

The language of theory building

$\mathcal{L}_{TB}$ is a five-tuple:

$$\mathcal{L}_{TB} = \langle \mathsf{lc}, \mathsf{con}, \mathsf{var}, \mathsf{term}, \mathsf{form} \rangle,$$

where lc stands for the set of logical constants:

$$\mathsf{lc} = \{(,), [,], \neg, \rightarrow, \forall, =, \mathfrak{N}, \mathfrak{P}, \mathfrak{A}\},$$

con represents the set of non-logical constants, falling into the two usual categories: predicates and individual constants, $\mathsf{con} = \mathsf{pred} \cup \mathsf{ind}$. The set of predicates is partitioned into (potentially) infinitely many subcategories according to the number of argument slots: $\mathsf{pred} = \cup_{n \in \omega} \mathsf{P}^n$, where P^n is the set of n-argument predicates and ω is the set of natural numbers. var

is the (infinite) set of variables. term, which stands for the set of terms, is defined as the union of ind and var. We assume that all of these sets are pair-wise disjoint.

Well-formed formulæ

The set of well-formed formulæ, form, is defined in the usual recursive manner. It is the smallest set that satisfies the following properties.

(1) A predicate filled up with the appropriate number of terms gives a formula. In formal terms:
if $a_1, \ldots, a_n \in \mathsf{term}$ and $P \in \mathsf{P}^n$, then $P(a_1, \ldots, a_n) \in \mathsf{form}$.

(2) The negation of a formula is a formula as well:
if $\phi \in \mathsf{form}$ then $\neg\phi \in \mathsf{form}$.

(3) The conditional between of two formulæ is a formula:
if $\phi \in \mathsf{form}$ and $\psi \in \mathsf{form}$, then $(\phi \rightarrow \psi) \in \mathsf{form}$.

(4) A formula prefixed with a universal quantifier is a formula:
if $x \in \mathsf{var}$ and $\phi \in \mathsf{form}$, then $\forall x[\phi] \in \mathsf{form}$.

(5) A *conditional* formula prefixed with any of the intensional quantifiers is a formula:
if $\bar{x} \subset \mathsf{var}$ and $\phi, \psi \in \mathsf{form}$, then $\mathfrak{N}\bar{x}[\phi \rightarrow \psi] \in \mathsf{form}$.
if $\bar{x} \subset \mathsf{var}$ and $\phi, \psi \in \mathsf{form}$, then $\mathfrak{P}\bar{x}[\phi \rightarrow \psi] \in \mathsf{form}$.
if $\bar{x} \subset \mathsf{var}$ and $\phi, \psi \in \mathsf{form}$, then $\mathfrak{A}\bar{x}[\phi \rightarrow \psi] \in \mathsf{form}$.

(6) The identity of two terms is a formula: if $a_1, a_2 \in \mathsf{term}$, then $\ulcorner a_1 = a_2 \urcorner \in \mathsf{form}$.

Semantics

We develop the semantics for our language in several steps. First we define a declarative semantics for the classical first-order fragment and the causal stories. The semantics of the causal stories is what Carlsson (1995) suggested: the causal stories are true if and only if the corresponding regularity is present in the model. (What we need to add is a construction that models the regularities.) Then we partialize this semantics to give a somewhat more realistic account on the information available in the context of theory building. We assume that some (classical) sentences are known to be true, some known to be false, and other sentences do not belong to either group yet. Similarly some of the causal stories are known, but it is unrealistic to assume that all relevant causal stories are known. These considerations lead to a construction where all of the classical sentences, causal stories, presumptions and auxiliary assumptions are valuated in sets of possible worlds, according to a set of causal stories. Such pairs made of sets of possible worlds and regularities capture what is known (in a given situation). These objects,

which we call scenarios, resemble the information states that play such a central role in the tradition of dynamic semantics developed by Kamp, Heim, Groenendijk, Stokhof, Veltman, and others. In this setup, classical sentences may have any of the classical truth-values or the value "unknown," while all non-classical sentences are either true or false.

Next we describe how theory stages define scenarios. This description resembles to the definition of update conditions for different types of sentences, except that our description is order-invariant. The dynamic semantics of the Amsterdam school was designed to represent sentences in a discourse; and the order of the sentences in a discourse obviously affects their meaning. So the order in which premises enter information states matters in these dynamic logics. However, we do not think that the order of the premises matter in our rendering of theory building. So our scenarios differ from information states in that they do not attend to the order of entry of premises.

Notation Let $U \neq \emptyset$ *denote the universe of discourse,* w *denote the set of possible worlds,* v *denote the set of all valuation functions, and* $\mathcal{P}$ *denote the powerset operation.*

The most convenient way to characterize an interpretation is to give the interpretation function.

Interpretation function

The interpretation function ρ, defined on the set of non-logical constants (con), satisfies the following conditions:

(1) for all $a \in \mathsf{ind}$, it is the case that $\rho(a) \in U$;
(2) for all $P \in \mathsf{P}^n$, $\rho(P) : \mathsf{w} \rightarrow \mathcal{P}(U^n)$.

It is clear that we need a formal representation of a regularity (or causal story). We argue that regularities should be represented as pairs of (open) formula intensions. We define the set of potential regularities, r, in two steps as follows.

First we deal with bare regularities, that is with regularities that do not embed other regularities.

Definition 1: (Set of bare regularities) *The set of bare regularities* (br) *is the set of ordered pairs of open formula intensions, i.e., pairs of mappings of possible worlds to variable valuations. The intuition behind this definition is that (1) regularities have a (sometimes implicit) if–then structure and both the "if" and the "then" parts are expressible with open formulæ, and (2) it is sufficient to know of these open formulæ which valuations of the variables*

make them true in which possible worlds. So let br *defined as follows:*

$$\mathsf{br} = \{f | f : \mathsf{w} \rightarrow \mathcal{P}(\mathsf{v})\}^2.$$

If we want a general definition of the set of regularities, we have to allow that regularities may embed other regularities. This possibility, in turn, means that the antecedent and the consequent parts of these complex regularities have somewhat more intricate notion of intensions. Possible worlds and variable valuations are not sufficient to tell whether the antecedent or the consequent is true or false. One must take into account the set of more primitive regularities too. To capture this intuition we offer an inductive definition of the set of regularities, where the induction operates on the levels of embedding. (r_1, the first level of regularities is basically the set of bare regularities, in a disguise.) To get all the regularities we take the union of all different levels of regularities.

Definition 2: (Set of regularities)

- $\mathsf{r}_0 = \emptyset$;
- $\mathsf{r}_n = \{f | f : \mathsf{w} \times \mathcal{P}(\mathsf{r}_{n-1}) \rightarrow \mathcal{P}(\mathsf{v})\}^2$;
- $\mathsf{r} = \cup_{n \in \omega} \mathsf{r}_n$.

If interpretation is given, then we can assign truth-values to all first-order formulæ according to one valuation or another. To work out the details of truth-value assignment in the case of regularities, we proceed it two steps. First we consider possible scenarios, which we defined above as pairs consisting of a possible world and a set of regularities.

We want to make clear that the truth-value assignments in scenarios do not provide the intended semantics for our purposes. Once we know for which possible world we should evaluate our formulæ, i.e., we know what exactly is the case, we should not bother with rules-with-exceptions. In other words, if all of the facts were known, then we would not need to look at regularities. But, we assume that it is never the case that all the facts are known in the case of real theories in flux.

The second semantics applies to *theory stages*, defined as pairs of sets of possible worlds and sets of regularities. A multiplicity of possible worlds represents the gaps in our knowledge. Our knowledge is generally partial. If we are working from a dependable body of knowledge, then we might be able to exclude certain possible worlds. But, if several worlds are still possible on the basis of our knowledge, then it makes sense to make use of weaker knowledge about regularities. If we do not know whether ϕ or $\neg\phi$ is the case in the real world, then we might sensibly base our scientific

predictions on default considerations, on regularities telling whether ϕ or $\neg\phi$ is *normally* the case.

A possible scenario is given by $\sigma = \langle w, \mathsf{r}' \rangle$, where w is a possible world i.e., $w \in \mathsf{w}$ and r' is a set of regularities, $\mathsf{r}' \subseteq \mathsf{r}$. The truth-value of the formula $\phi \rightarrow \psi$ in this scenario according to the valuation v is denoted by $\|\phi \rightarrow \psi\|_v^\sigma$.

The definition of truth-values in scenarios (according to valuations) follows the recursive definition of formulæ. We define truth-values and open-formula intensions in parallel.

Truth definitions in scenarios

Let σ be a scenario and v a valuation.

(1) $\|a\|_v^\sigma = (\rho \cup v)(a)$ (If a is an individual constant, then it is interpreted by ρ; if it is a variable, then it is valuated by v.) It is easy to see that the intensions of terms are independent of the possible worlds and the set of regularities.

(2) $\|P\|_v^\sigma = \rho(P)(w)$ (If P is a predicate, then it is interpreted by ρ; this is its intension. However, the extension of the predicate might differ among possible worlds. No dependence on the set of regularities or on valuations.

(3) $\|a = b\|_v^\sigma = 1$ if $\|a\|_v^\sigma = \|b\|_v^\sigma$, and $\|a = b\|_v^\sigma = 0$ otherwise. Identity statements depend only on the denotation of the terms; they are independent of possible worlds and sets of regularities.

(4) $\|P(a_1, \ldots, a_n)\|_v^\sigma = 1$ if $\langle \|a_1\|_v^\sigma, \ldots, \|a_n\|_v^\sigma \rangle \in \|P\|_v^\sigma$, and $\|P(a_1, \ldots, a_n)\|_v^\sigma = 0$ otherwise.

(5) $\|\neg\phi\|_v^\sigma = 1$ iff $\|\phi\|_v^\sigma = 0$.

(6) $\|\phi \rightarrow \psi\|_v^\sigma = 0$ if $\|\phi\|_v^\sigma = 1$ and $\|\psi\|_v^\sigma = 0$; and $\|\phi \rightarrow \psi\|_v^\sigma = 1$ otherwise.

(7) $\|\forall x[\phi]\|_v^\sigma = 0$ if there is an $a \in U$, such that $\|\phi\|_{v[x:a]}^\sigma = 0$; and $\|\phi\|_{v[x:a]}^\sigma = 1$ otherwise.

(8) $\|\phi\| : \mathsf{w} \times \mathcal{P}(\mathsf{r}) \rightarrow \mathcal{P}(\mathsf{v})$ such that $v \in \|\phi\|(w, \mathsf{r}')$ if and only if $\|\phi\|_v^{\langle w, \mathsf{r}' \rangle} = 1$.

(9) $\|\mathfrak{N}\bar{x}[\phi \rightarrow \psi]\|_v^{\langle w, \mathsf{r}' \rangle} = 1$ if $\langle \|\phi\|_{v|_{\bar{x}}}^{\langle w, \mathsf{r}' \rangle}, \|\psi\|_{v|_{\bar{x}}}^{\langle w, \mathsf{r}' \rangle} \rangle \in \mathsf{r}'$; and $\|\mathfrak{N}\bar{x}[\phi \rightarrow \psi]\|_v^{\langle w, \mathsf{r}' \rangle} = 0$ otherwise.

We do not define the truth-values of formulæ prefixed with the $\mathfrak{P}$ or $\mathfrak{A}$ quantifiers in scenarios. It only makes sense to define them in theory stages. In scenarios, where the truth-value of any (first-order) formulæ is "known," presumptions are useless, and at best, misleading, as we noted above.

Now we move on to define theory stages and a semantics for all our formulae in theory stages.

Preparation for the Theory-Stage Semantics. A theory stage is a formal rendering of the incomplete information provided by a theory in flux. There are two sources of partiality. The first involves gaps in factual knowledge. It might be the case that some facts, expressible as first-order formulæ, are known to hold, that other formulæ are known not to hold, and some are neither known to hold nor to not hold. This being the case, a theory stage has to provide a partial semantics, a semantics that makes some classical first-order formulæ true, some others false, and allows a truth-value gap for the rest.

In terms of a possible worlds, knowing the truth-values of all classical formulæ means knowing exactly which of the possible worlds is the actual one. Knowing only some of the truth-values of the non-tautological sentences means knowing only that a subset of the set of possible worlds includes the actual world, but not knowing exactly which of the candidates is the actual world.

Due to this motivation, it appears to be natural that one component of a theory stage is a set of worlds that are still possible (possible given the state of knowledge at the stage of the theory). An expansion of a theory (development of a new stage) might eliminate some of the worlds that were still possible in the previous state. No theory expansion will bring back any of the already-eliminated possible worlds. But, there can be theory extensions that leave the set of still-possible worlds intact. These are the ones that operate on the second source of partiality.

The second source of partiality is related to the first. Although an empirical theory cannot provide complete information about the facts, it might still be capable of providing regularities, which can be used to fill some of the gaps in knowledge. But, as it happens, the set of regularities that a stage of a theory can provide might also be incomplete. In case of theories in flux, this information is indeed incomplete. Observations and thought experiments typically bring some regularities in sight but fail to provide all that are needed. Therefore a theory stage will be equipped a subset of regularities and will support a set of causal stories. Theory extensions sometimes take the form of incorporating some new regularity. Here again theory extensions do not eliminate established regularities, but they occasionally add new ones.

This picture suggests that the process of theory building is monotonic: information is added but never withdrawn. Whatever belongs to a theory stage is persistent, it belong to all extensions of that stage. Instead of being withdrawn, regularities might be partially — or even completely — overridden by some more-specific regularity.

We conclude these considerations with definitions of a theory in flux and of a theory stage.

Definition 3: (Theory in flux) *Let π_f be a set of facts (premises expressed as sentences of LoFoL), and π_n be a set of formulae expressing rules-with-exceptions, i.e., formulæ prefixed by the $\mathfrak{N}$ intensional quantifiers. We refer to the $\langle \pi_f, \pi_n \rangle$ pair as a theory in flux.*

Definition 4: (Theory stage) *The pair $\langle \mathsf{w}', \mathsf{r}' \rangle$ is a theory stage if*

(1) $\mathsf{w}' \subseteq \mathsf{w}$, $\mathsf{r}' \subseteq \mathsf{r}$, *and*

(2) *if $\langle a, b \rangle \in \mathsf{r}'$ and $\langle c, b \rangle \in \mathsf{r}'$, then $\langle d, b \rangle \in \mathsf{r}'$, where*

$$d : \mathsf{w} \times \mathcal{P}(\mathsf{r}) \rightarrow \mathcal{P}(\mathsf{v}), \text{ and } v \in d(w, \mathsf{r}') \leftrightarrow v \in a(w, \mathsf{r}') \vee v \in c(w, \mathsf{r}').$$

Because causal stories lack rich, "fully compositional" semantics, we leave it to the semantics of the formulæ prefixed with $\mathfrak{P}$ and $\mathfrak{A}$ to characterize the basic intuitions about how we make inferences from rules with exceptions. Intuitively $\mathfrak{P}[\phi \rightarrow \psi]$ is true in a theory stage if one can construct a tentative, but convincing, argument based on the established facts and the available set of regularities, and at least one of such tentative arguments is more specific than all the (tentative) counter-arguments. Tentative arguments are going to be represented by chains of regularities. Regularities in these chains are semantic representations of rules (strict rules, rules-with-exceptions, definitions, and meta-considerations). The semantic rendering of a theory in flux represents rules in the form of pairs of formula intensions, the first component of which is the intension of the antecedent, and the second component is the consequent.

We have to define the proper construction of the chain: which regularity can follow which other regularity in the chain. As a preparatory step, we need a transitive and reflexive relation on the set of formula intensions that captures the notion of *degree of specificity*. The specificity ordering of a pair of regularities can be characterized as follows. In all still-possible worlds (given a theory stage) the intension of the antecedent of one regularity is smaller than or equal to the intension of the antecedent of the other regularity.

Definition 5: (The specificity relation for formula intensions) *Let $\mathsf{w}' \subseteq \mathsf{w}$, $\mathsf{r}' \subseteq \mathsf{r}$, and let a and b each be elements of $\{x | x : \mathsf{w} \times \mathcal{P}(\mathsf{r}) \rightarrow \mathsf{v}\}$. a is more specific than b, ($a \sqsubseteq_{\mathsf{w}',\mathsf{r}'} b$)), iff for all $w \in \mathsf{w}'$ it holds that $a(w, \mathsf{r}') \subseteq b(w, \mathsf{r}')$.*

With this relation in hand, we can define regularity chains. Each component in the chain is a regularity; hence it is given by a pair of open-formula intensions. We use the following notation in referring to the subcomponents of an element in a chain. Let ρ_i^1 denote the first element (antecedent) in the pair of intensions that compose the regularity that sits in the ith position in

the chain, ρ_i^2 denote the second element in that regularity (the consequent) in the link.

Definition 6: (Regularity chain) *Let* $\langle \mathsf{w}', \mathsf{r}' \rangle$ *be a theory stage.* ($\mathsf{w}' \subseteq \mathsf{w}$ *and* $\mathsf{r}' \subseteq \mathsf{r}$.) *A sequence* $\langle \rho_0, \rho_1, \ldots, \rho_k \rangle$ *is a k-step* $\phi \rightarrow \psi$ *positive regularity chain (or alternatively a k-step* $\phi \rightarrow \neg\psi$ *negative regularity chain)* $\langle \mathsf{w}', \mathsf{r}' \rangle$ *if:*

(1) $\rho_0^1 \sqsubseteq_{\langle \mathsf{w}', \mathsf{r}' \rangle} \rho_0^2$
(2) $\rho_0^1 = \|\phi\|$ *and* $\rho_k^2 = \|\psi\|$;
(3) $\forall i[1 \leq i \leq k \rightarrow \rho_i \in \mathsf{r}']$;
(4) $\forall i[1 < i \leq k \rightarrow \rho_{i-1}^2 \sqsubseteq_{\langle \mathsf{w}', \mathsf{r}' \rangle} \rho_i^1]$[4].

We call $\rho_1 = \langle \rho_1^1, \rho_1^2 \rangle$ *the initial link in the regularity chain.*

Definition 7: (Specificity ordering of the regularity chains) *Let* $\langle \mathsf{w}', \mathsf{r}' \rangle$ *be a theory stage. The regularity chain* $\langle \rho_1, \ldots, \rho_k \rangle$ *(in* $\langle \mathsf{w}', \mathsf{r}' \rangle$*) is* more specific *than the regularity chain* $\langle \rho_1', \ldots, \rho_l' \rangle$ *(also in* $\langle \mathsf{w}', \mathsf{r}' \rangle$*) if* $\rho_i^1 \sqsubseteq_{\langle \mathsf{w}', \mathsf{r}' \rangle} \rho_j'^1$ *and* $\rho_j'^1 \not\sqsubseteq_{\langle \mathsf{w}', \mathsf{r}' \rangle} \rho_i^1$, *or if* $\langle \rho_0, \rho_1, \ldots, \rho_k \rangle$ *has no initial element.*

According to this definition only the antecedent of the initial step in the chain determines the specificity of the chain. Figure 1 shows the motivation for this choice. (The ellipses in Figure 1 indicate the relationships between formula intensions.) It illustrates that, even when we deliberately try to construct a case where the specificity order on the first link of the rule chains works opposite to that the second link, the first of these picks out the more specific argument.

Now we can give the semantics for the formulae prefixed with the intensional quantifiers $\mathfrak{P}$, and $\mathfrak{A}$.

Definition 8: (Semantics of presumptions and auxiliary assumptions)

- $\|\mathfrak{P}\bar{x}[\phi \rightarrow \psi]\|_v^{\langle \mathsf{w}', \mathsf{r}' \rangle} = 1$ *if:*
  (1) *there exists a positive, one element* $\phi \rightarrow \psi$ *regularity chain in* $\langle \mathsf{w}', \mathsf{r}' \rangle$ *or*
  (2) *there exist positive* $\phi \rightarrow \psi$ *regularity chains (of length two or longer) in* $\langle \mathsf{w}', \mathsf{r}' \rangle$ *and if there also exist negative* $\phi \rightarrow \psi$ *regularity chains (of length two or longer) in* $\langle \mathsf{w}', \mathsf{r}' \rangle$, *then at least one positive regularity chain is more specific than any negative regularity chain.*

[4] The intension of an open formula, depends on the order of free variables. Therefore, we would actually require that $\langle \rho_i^1 \rangle, \pi(\rho_i^2) \rangle \in \mathsf{r}'$ for some permutation π. We will ignore this technicality in the rest of the paper.

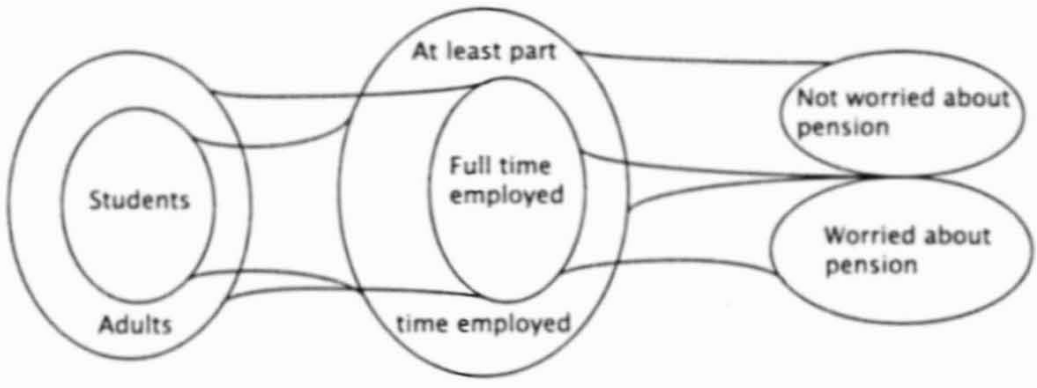

Figure 1. The specificity of the argument depends on its first premise only

$\|\mathfrak{P}\bar{x}[\phi]\|_v^{\langle \mathsf{W}', \mathsf{r}' \rangle} = 0$ *otherwise.*

- $\|\mathfrak{A}\bar{x}[\phi]\|_v^{\langle \mathsf{W}', \mathsf{r}' \rangle} = \|\mathfrak{P}\bar{x}[\phi]\|_v^{\langle \mathsf{W}', \mathsf{r}' \rangle}$.

Inferencing within Theories in Flux. The inferencing we are interested in starts with premises expressed in first-order logic and in terms of rules-with-exceptions and auxiliary assumptions; and conclusions that are expressed as presumptions. Now we define the semantic consequence relation as follows:

Definition 9: (Stages of a theory in flux) *The stage of the theory that correspond to* $\langle \pi_f, \pi_n, \rangle$ *is the pair of a set of possible worlds and a set of regularities,* $\langle \mathsf{W}', \mathsf{r}' \rangle$, *if it is a theory stage and the following conditions are met:*

(1) $\mathsf{W}' = \{w \in \mathsf{W} | \forall \pi [\pi \in \pi_f \rightarrow \|\pi\|_v^{\langle w, \mathsf{r}' \rangle} = 1]\}$.

(2) $\forall \pi [\pi \in \pi_n \rightarrow \|\pi\|_v^{\langle w, \mathsf{r}' \rangle} = 1]\}$;

(3) *If* r'' *satisfies 1, and 2, then* $\mathsf{r}' \subseteq \mathsf{r}''$.

The final condition guarantees that the regularity set is the smallest that the theory in flux requires. (We do not want to assume more regularities than the theory in flux requires.)

Definition 10: (Theory stage with auxiliary augmentations) *Let π_a be a set of auxiliary assumptions. The stage of the theory that corresponds to $\langle \pi_f, \pi_n, \pi_a \rangle$ is $\langle \mathsf{W}', \mathsf{r}', \mathsf{r}'' \rangle$ if the following conditions are met:*

(1) *$\langle \mathsf{W}', \mathsf{r}' \rangle$ is the theory stage for $\langle \pi_f, \pi_n \rangle$*
(2) *if $\langle \mathsf{W}', \mathsf{r}' \rangle$ is the theory stage.*
(3) *$\mathsf{r}' \subseteq \mathsf{r}''$, and the following conditions are met*
- *if $\mathfrak{A}\bar{x}[\phi \rightarrow \psi] \in \pi_a$ and ϕ' is such that $\phi' \sqsubseteq_{\langle \mathsf{W}', \mathsf{r}' \rangle} \phi$, and $\|\mathfrak{P}\bar{x}[\phi' \rightarrow \psi]\|_v^{\langle \mathsf{W}', \mathsf{r}' \rangle} = 1$ but in case $\phi' \sqsubseteq_{\langle \mathsf{W}', \mathsf{r}' \rangle} \phi'' \sqsubseteq_{\langle \mathsf{W}', \mathsf{r}' \rangle} \phi$ it follows that $\|\mathfrak{P}\bar{x}[\phi'' \rightarrow \psi]\|_v^{\langle \mathsf{W}', \mathsf{r}' \rangle} = 0$ then $\|\mathfrak{N}\bar{x}[\phi \wedge \neg\phi' \rightarrow \psi]\| \in \mathsf{r}_v^{\langle \mathsf{W}', \mathsf{r}'' \rangle} = 1$*
- *r'' is the smallest set that satisfies the conditions above.*

Definition 11: (Implications of theories in flux) *Let ϕ be a formula. $\pi_f \cup \pi_n \cup \pi_a$ logically implies ϕ iff the corresponding stage of the augmented theory $\pi_f \cup \pi_n \cup \pi_a$ makes ϕ true too.*

As it obvious from Figure 2, the augmentation with very same auxiliary assumption has a different impact in different theory stages. Is makes the most specific change necessary to guarantee the truth of the auxiliary statement. These assumptions remain true with theory expansions, but their impact is not at all constant. We believe that the lack of constant impact on the model is a formal confirmation that these assumptions do not belong to the core of the theory. We offer an additional characteristic of these assumptions below that point in the same direction.

Falsifiability

So far we only claimed informally that the auxiliary assumptions are not claims of the theory and therefore it does not make sense to take them as part of the core. Now we are in the position to bring this issue one step further and offer a formal characterization of empirical testability.

Definition 12: *Let $\langle \pi_f, \pi_n, \pi_a \rangle$ be an augmented theory in flux, i.e $\langle \pi_f, \pi_n \rangle$ is a theory in flux and boldsymbolπ_a is the set of auxiliary assumptions we augment it with. Let furthermore $\mathfrak{Q}$ be one of the following quantifiers. $\forall, \mathfrak{N}, \mathfrak{A}$. A formula of the form $\mathfrak{Q}x[\phi(x) \rightarrow \psi(x)]$ is falsifiable in $\langle \pi_f, \pi_n, \pi_a \rangle$ $\mathfrak{Q}x[\phi(x) \rightarrow \psi(x)]$ is true in the corresponding theory stage*

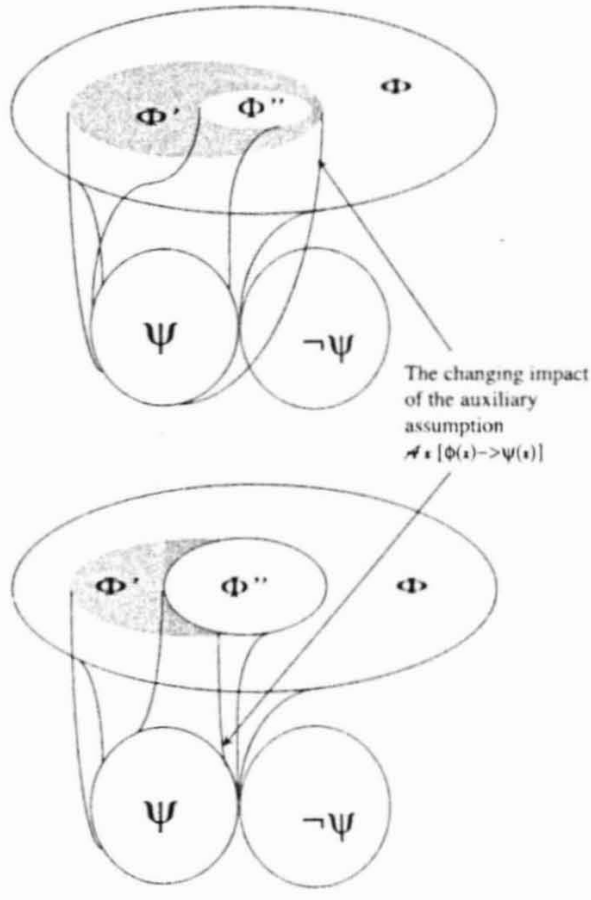

Figure 2. The shadowed part indicates where the auxiliary assumption has an impact.

$\langle \mathsf{w}', \mathsf{r}', \mathsf{r}'' \rangle$ *but there is a* $\langle \mathsf{w}'_1, \mathsf{r}'_1, \mathsf{r}''_1 \rangle$ *theory stage corresponding to some extensions of* $\langle \pi_f, \pi_n, \pi_a \rangle$ *which makes* $\mathfrak{P}x[\phi(x) \rightarrow \neg\psi(x)]$ *true.*

It is easy to see that neither the universally quantified formulæ nor the auxiliary assumptions are falsifiable in a theory fragment augmented with auxiliary assumptions according to this definition. For the auxiliary assumptions it is obvious because the corresponding theory stage with augmentations include the regularity that is just specific enough to make $\mathfrak{P}x[\phi(x) \rightarrow \psi(x)]$ true. All strict (classical first-order) arguments reduce to 0-step regularity chains in all stages of a theory. If the argument is sound in classical first-order logic, then there is a $\sqsubseteq_{\mathsf{w}',\mathsf{r}'}$ relation between the first and the last element of the chain for any w', r' theory stage. Such arguments are by definition more specific than any argument that appeals to regularities. As a result, first-order arguments overrule arguments based on regularities. Due to this feature of sound classical first-order arguments definitions and meta-considerations of a theory are never falsifiable.

Once the notion of the logical consequence relation is defined, our task of defining a logic is completed — in a way. Judging how well the logic fits our motivations requires more than establishing that all of the initial considerations are honestly implemented. It is also important that some properties of classical first-order logic that yield damagingly counter-intuitive results can no longer be reproduced in this logic. In particular, it is important to note that two classical inference rules do not hold: modus tollens and contraposition. We wanted to rule these operations out for the logic of theory building because we think that they require more dependable knowledge (about individual cases) than can be delivered by rules-with-exceptions. Insightful causal stories often expressed in a *ceteris paribus* form. Even when they are not expressed in this way, to interpret them as comparative statements made on the all-other-things-being-equal basis is the most defendable interpretation. Consider two sentences from a theory we developed with Glenn Carroll (Hannan, Pólos, and Carroll 2003b): "A more intricate organization has higher inertia" and "A more opaque organization has higher inertia."

If these sentences are formalized in the language of classical first-order logic, the pair of formal counterparts implies the conclusion "A more intricate is an organization has higher opacity." This last sentence might or might not be the case. But it appears that the implication relation does not hold among the informal (generic) sentences.

If we try to formalize these sentences in the logic of theory building, the translations look like these:

"Normally a more intricate organization has higher inertia" and "Normally a more opaque organization has higher inertia." But now the first two sentences do *not* imply: " More intricate organizations presumably have higher opacity" simply because we ruled out the contraposition operation in the new logic on the grounds that there is nothing in the set of regularities that would support such a conclusion.

Appendix: Theory unification

In what follows we show how the logic for theory building helps theory unification. We start with three theory fragments that belong to the same scientific research program, organization ecology, and we address the very same research question: What is the relationship between the age of organizations and their hazards of mortality. Historically these theory fragments have been developed to explain the empirical findings of studies carried out in different populations of organizations. These findings were contradictory: in some populations the mortality hazard appeared to be decreasing with age (liability of newness), in other populations the mortality hazard increased with age

(liability of obsolescence), and there were populations where the relationship was nonmonotonic, first the mortality hazard increased to a maximum but decreased after that (liability of adolescence). The need for theory unification was imminent. Without the unification the research program was unable to provide suggestions how the age dependence of the hazard of mortality might look like in a not yet studied population of organizations. Hannan (1999) showed that a formalization of these theory fragments in classical first-order logic can be used to unify the liability of newness and the liability of adolescence, but he concluded that all three of theory fragments could not be unified. Now we show briefly that the unification is both possible and insightful if the theory fragments are formalized in the nonmonotonic logic for theory building. A more detailed unification of these and few other theory fragments can be found in Pólos and Hannan (2000, 2002).

We start with some notation. Let $\mathrm{O}(o,p)$ be a predicate that tells that o is member of organizational population p. Many of the assumptions and theorems in these theory fragments involve monotonicity statements. We simplify presentation of formulæ stating such relations by adopting notational shorthand. Suppose f is a function defined for organizations at time points. We usually denote such functions in the following format: $f(o,s)$, where o refers to an organization and s is a time point. We will often want to compare the values of these functions for different organizations (in the same population) and time points. We use the expression $f \uparrow g$ to indicate a monotonic positive relationship between the two functions, and $f \downarrow g$ to indicate a monotonic negative relationship[5].

Fragment 1: Liability of newness

We formalize the argument about age-related capabilities using the nonnegative function, $cap(o,s)$, that records o's level of capability at the time s. We continue to represent organization o's age at time s with the nonnegative function $a(o,s)$. Since the all the "normally" and the "presumably" quantifiers uniformly range over the same five variables o, o', p, s, s' and these variables remain implicit in the above introduced notation for monotonicity statements we use the shorthand $\mathfrak{N}$ for $\mathfrak{N}o, o', p, s, s'$ and $\mathfrak{P}$ for $\mathfrak{P}o, o', p, s, s'$ respectively.

Postulate 1: *An organization's expected level of capability increases with its age.*

$$\mathfrak{N}[\,a \uparrow \mathrm{E}\{cap\}]^{6}.$$

[5] A more precise definition of $f \downarrow g$ can be found in Pólos and Hannan (2002).

Postulate 2: *Higher capability lowers the mortality hazard.*

$$\mathfrak{N}[\, cap \downarrow \mu].$$

Now we want to connect these two postulates, which form a chain — except that the consequent in the first postulate is a comparison of expected levels of capability and the antecedent in the second postulate contains a comparison of the actual levels of capability. Hannan, Pólos and Carroll (2005) argue for a metarule that allows such formula to be chained. Using this metarule, we have the strong-form version of the liability of newness:

Theorem 1: *Mortality hazards decline monotonically with age.*

$$\mathfrak{P}[a \downarrow \mu].$$

The ellipses in Figure 3 again feature the open-formula intensions while the shapes between them are representations of the explanatory principles that connect them.

We treat this first stage as the default theory. Its postulates will be included in every subsequent stage. Notice that, because this (provisional) theorem applies to any age interval, its scope of applicability is extremely *non-specific*. It will turn out that more specific postulates in the more developed theory fragments usually override it over at least part of the age range.

Fragment 2: Endowments

The next development introduced endowments. An organization is founded with a given level of endowment if it possesses immunity after founding, at least for a time. Endowment lasts as long as this initial immunity does. Furthermore, there is a monotonic relation between the level of endowments and the strength of immunity. The higher the level of endowment the stronger the immunity. Let $ed(o, s)$ tell the level of endowment of organization o at age s and let $im(o, s)$ give the level of immunity.

[6] This formula reads as follows: It is normally the case for all pairs of organizations (in a population) at all pairs of ages that the expected level of capability at an older age exceeds that at a younger age.

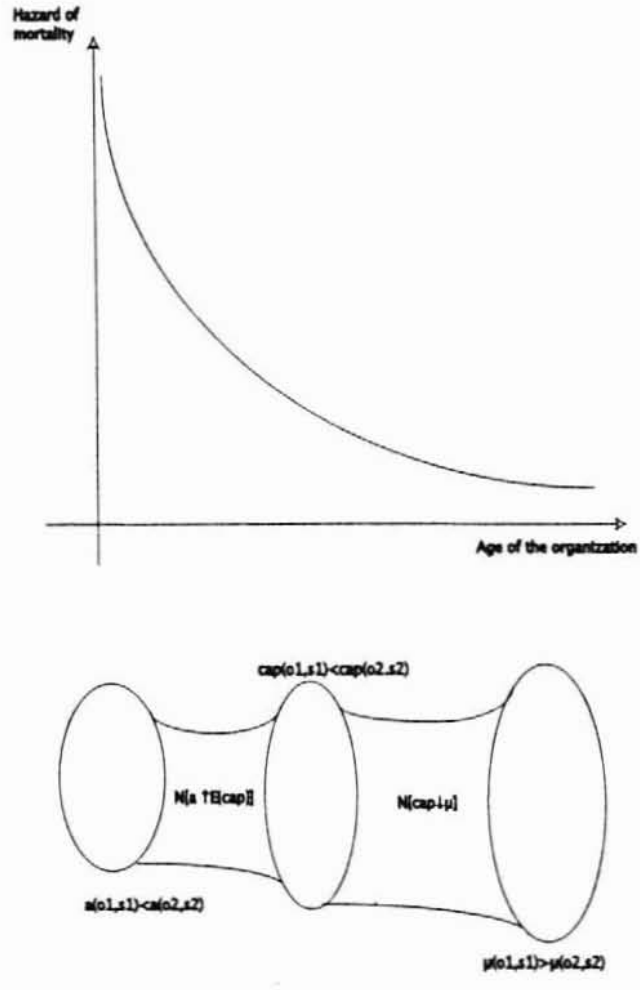

Figure 3. The graphic representation of Theorem 1 and its derivation

Auxiliary assumption 1: *The expected age at the ending of initial endowm is constant within an organizational population.*

$$\mathfrak{A}p[\mathrm{P}(p) \rightarrow \exists \zeta_p \forall o, o', s, s'[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \rightarrow$$
$$\sup\{t \mid (a_o(s) = t) \wedge \mathrm{E}\{ed_o(s)\} > 0\} = \zeta_p$$
$$= \sup\{t' \mid (a_{o'}(s') = u') \wedge \mathrm{E}\{ed_{o'}(s')\} > 0]$$

Note that this auxiliary assumption instantiates the premise that the expec age of ending of endowment is the same for all members of a population a also labels this expectation as ζ_p. Henceforth, we let ζ_p denote the expec ending time of endowment for population p.
The standard argument holds that organizations normally spend down th initial endowments.

Postulate 3: *Expected levels of endowments decline monotonically within endowment periods.*

$$\mathfrak{N}[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \wedge (a(o,s) < a(o',s') < \zeta_p) \rightarrow \mathrm{E}\{ed(o,s)\} > \mathrm{E}\{ed(o',s')\}].$$

Moreover, endowments provide immunity and immunity brings a reduction in mortality chances. These postulates hold both for comparisons of an organization at different ages (say before and after the ending of endowment) and for pairs of organizations (say, with different levels of immunity).

Postulate 4: *During a period of endowment, a larger endowment yields a higher expected level of immunity.*

$$\mathfrak{N}[ed \uparrow \mathrm{E}\{im\}].$$

Postulate 5: *Mortality hazards fall with increasing immunity.*

$$\mathfrak{N}[im \downarrow \mu].$$

Theorem 2: *Mortality hazards increase with age within endowment periods.*

$$\mathfrak{P}[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \wedge (a_o(s) < a_{o'}(s') < \zeta_p) \rightarrow \mu_o(s) < \mu_{o'}(s')].$$

Theorem 3: *Mortality hazards are lower within endowment periods than afterwards.*

$$\mathfrak{P}[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \wedge (a_o(s) < \zeta_p \leq a_{o'}(s')) \rightarrow \mu_o(s) < \mu_{o'}(s')].$$

The First Unification Attempt

A key step in developing a modeling procedure involves translating the verbal argument into a formal language that enables nonmonotonic testing. It is easy to realize that the claim "Endowment considerations apply only before the end of the endowment period" is not specific enough. Even though it makes clear that the considerations are not applicable to intervals beginning after the endowment is exhausted they might or might not be applicable to intervals that begin before and finish after the end of the endowed period. Both possible translations (that the considerations apply and that they do

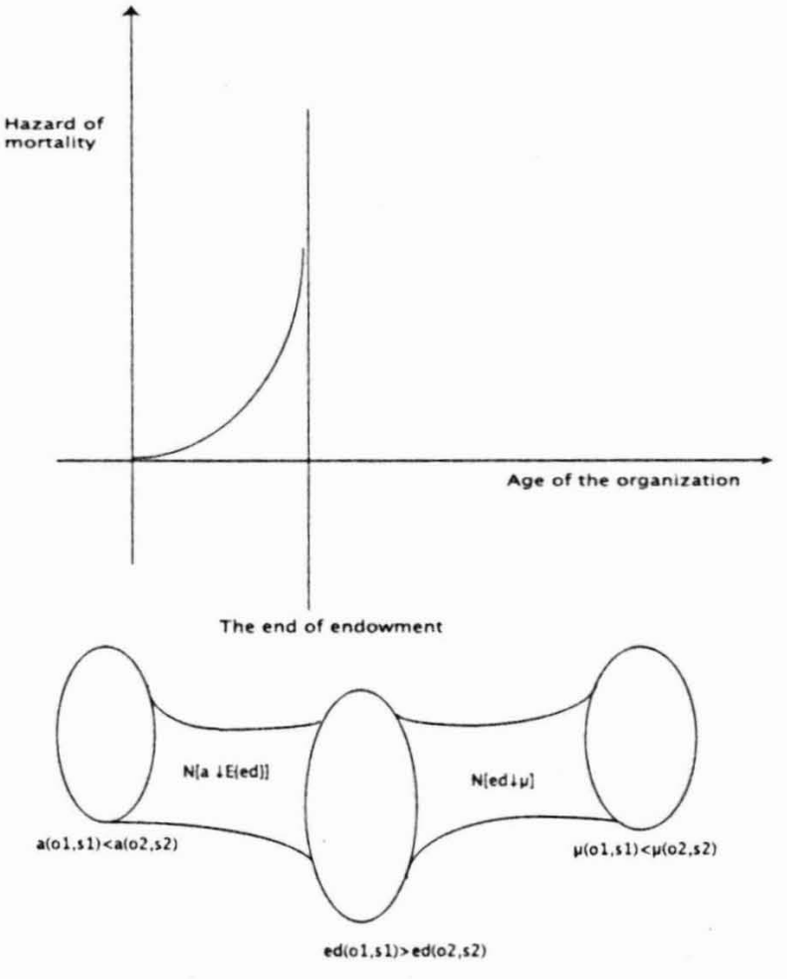

Figure 4. Age dependence of mortality hazard during the endowed period, and the regularity chain that supports it.

not apply to this type of intervals) are consistent with the nonmonotonic approach; and in both cases we see a penguin scenario. Still, one of them might be more in line with the concept of the mortality hazard than the other.

Let us consider first the option: endowment considerations do not apply to intervals that stretch over the end of the endowment. Under this restriction,

only one line of argument applies: the default theory of a liability of newness. According to this theory, the hazard at the beginning of the interval exceeds the hazard at the end of the interval. Due to the immunity considerations, the hazard at the very beginning of an (endowed) organization's life is zero, according to the first translation of the key claim. Now take an interval that begins immediately after the founding of the organization and ends some time after the end of the endowment period. At the end of such a period, the hazard must be *negative.* Although this scenario is a logical possibility, it violates the definition of a hazard.[7] It is tempting to look at this conclusion as a case for a non-theorem in the desiderate of the research program. Therefore, this approach does not meet the most basic requirement for a modeling procedure for mortality processes. The second translation holds that endowment considerations do apply to this type of intervals. This translation does not generate the undesirable result of implying negative hazards. Moreover, we will show that it yields interesting results.

The first unification attempt uses all four postulates in the two fragments (according to the strategy we outlined) to yield:

Theorem 4: *Mortality hazards increase with age over intervals that begin within expected endowment periods, that is, before* ζ_p.

$$\mathfrak{P}[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \wedge (a_o(s) < a_{o'}(s') < \zeta_p) \rightarrow \mu_o(s) < \mu_{o'}(s')].$$

Proof. Figure 5 depicts the relevant regularity chains. One begins with the intension defined for any pair of ages (drawn as the large ellipse at the top of the figure.) This rule chain leads to the conclusion of negative age dependence. The regularity chain drawn on the left emanates from the smaller (more specific) intension that applies only to those age intervals that begin before the expected ending of endowment. This regularity chain leads to a conclusion of positive age dependence. According to the nonmonotonic inference rule, the more specific argument holds.[8] □

[7] The hazard is defined as the ratio of two non-negative functions, the density of the ending durations and the survivor function. Therefore, a negative value of a hazard entails a contradiction.

[8] It might seem from this example that the less specific regularity chain should dominate because it is shorter. However, this is not the case. Length of chains matters for testing only when the chains being compared have the same specificity.

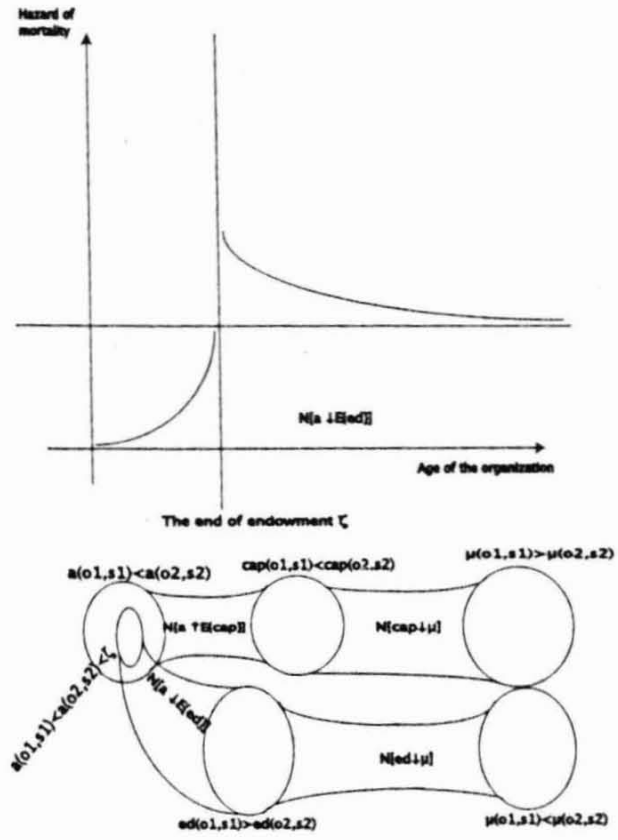

Figure 5. The more specific argument dominates. Picture of the first unification attempt and its derivation

Theorem 5: *Mortality hazards decrease with age after endowments are exhausted.*

$$\mathfrak{P}[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \wedge (\zeta_p \leq a_o(s) < a_{o'}(s')) \rightarrow \mu_o(s) > \mu_{o'}(s')].$$

Proof. Examine the most-specific regularity chains that connect the intension of $\zeta_p < a_o(s) < a_{o'}(s')$ with the intension of $\mu_o(s) > \mu_{o'}(s')$. The only rule chain that applies is the less specific one (on the right of the figure) that leads to the conclusion of negative age dependence. □

A corollary also follows from these two theorems: an overall tendency toward *positive* age dependence, as sketched in Figure 5. We regard this result as somewhat surprising in the sense that organizational theorists, in focusing on the different fragments, did not notice this implication. We shared this limited vision when we set out to construct a model, and we were pleasantly surprised to learn that the postulates as formulated in nonmonotonic logic delivered more than we had expected. The effort to unify fragments in a consistent manner (at the same level of analysis) makes clear the importance of these subtle differences in assumptions.

Theorem 6: *An organization's mortality hazard jumps to its maximum when its endowment ends.*

Fragment 3: Obsolescence

Now we turn to the other main branch of the theory, which concerns positive age dependence. We concentrate on the version that relies on assumptions about obsolescence. We assume that the quality of the alignment between organizations and their environments affects mortality chances. We also assume, that organizations are relatively inert and, in the long run, their structures cannot follow environmental drift. The drift is such that, within a period of length ω_p, the quality of alignment for organization o does not change so much from the founding conditions that it affects the hazard. However, beyond ω_p, the environment has normally drifted far enough as to drive the quality of alignment below a threshold that affects the hazard. Further drift, beyond ω_p, continually degrades alignment.

We introduce the non-negative function $al(o, s)$ that gives the level of alignment of organization o with its environment at s.

Environmental change drives the obsolescence process. Suppose that the environment can occupy different states at different times, in the sense that it imposes different adaptive demands at different times. Two states of an environment impose dissimilar adaptive demands if an organization cannot be aligned with both. Organization-builders can use state-of-the art designs and adapt to prevailing cultural understandings. This motivates the following auxiliary assumption

Auxiliary assumption 2: *Organizations have nonzero (expected) alignment with their environments at founding.*

$$\mathfrak{A}o, p, s[\mathrm{O}(o, p) \wedge (a(o, s) = 0) \rightarrow \mathrm{E}\{al_o(s)\} > 0].$$

Definition 13: *Drifting environment for an organizational population*

$$\mathrm{DRIFT}(p) \leftrightarrow \mathfrak{N}o, s[\mathrm{O}(o, p) \wedge (a_o(s) > \omega_p) \rightarrow \mathrm{E}\{al_o(s)\} = 0].$$

Auxiliary assumption 3: *The expected age of obsolescence (ending of alignment) in an organizational population in a drifting environment is a constant.*

$$\mathfrak{A}p[\mathrm{P}(p) \rightarrow \exists\, \omega_p \forall o, o', s, s'[\mathrm{DRIFT}(p) \wedge \mathrm{O}(o, p) \wedge \mathrm{O}(o', p) \rightarrow \sup\{s \mid \mathrm{E}\{al_o(s)\} > 0\} = \omega_p = \sup\{s' \mid \mathrm{E}\{al_{o'}(s')\} > 0\}]].$$

Once organizations lose alignment with their environments, they start to become devalued by relevant evaluators as "obsolete." The longer an organization has been obsolete, the stronger is this devaluation process. Let $ob(o, s)$ be a function that tells the degree to which organization o's relevant audiences regard it as obsolete at age s.

Postulate 6: *After the onset of obsolescence, organizations are normally judged to be increasingly obsolete with further aging in drifting environments.*

$$\mathfrak{N}[\text{DRIFT}(p) \wedge \text{O}(o,p) \wedge \text{O}(o',p) \wedge (\omega_p \leq a_o(s) < a_{o'}(s')) \rightarrow \\ \text{E}\{ob_o(s)\} < \text{E}\{ob_{o'}(s')\}].$$

Postulate 7: *Higher perceived obsolescence yields higher mortality hazards.*

$$\mathfrak{N}[\, ob \uparrow \mu].$$

These premises imply a pair of theorems.

Theorem 7: *Mortality hazards are higher after the expected age of onset of obsolescence than before.*

$$\mathfrak{P}[\text{DRIFT}(p) \wedge \text{O}(o,p) \wedge \text{O}(o',p) \wedge (a_o(s) < \omega_p \leq a_{o'}(s')) \rightarrow \\ \mu_o(s) < \mu_{o'}(s')].$$

Theorem 8: *Mortality hazards increase with age after the expected age of onset of obsolescence.*

$$\mathfrak{P}[\text{DRIFT}(p) \wedge \text{O}(o,p) \wedge \text{O}(o',p) \wedge (\omega_p \leq a_o(s) < a_{o'}(s')) \rightarrow \\ \mu_o(s) < \mu_{o'}(s')].$$

The Second Unification Attempt

The second unification uses all of the definitions and postulates in the three theory fragments. Again we confront the issue of what to do with intervals for which a specific rule applies to part but not all (and, by definition, the default applies to the whole interval). Again, to avoid having the specific rule made irrelevant, we posit that whenever the more specific obsolescence

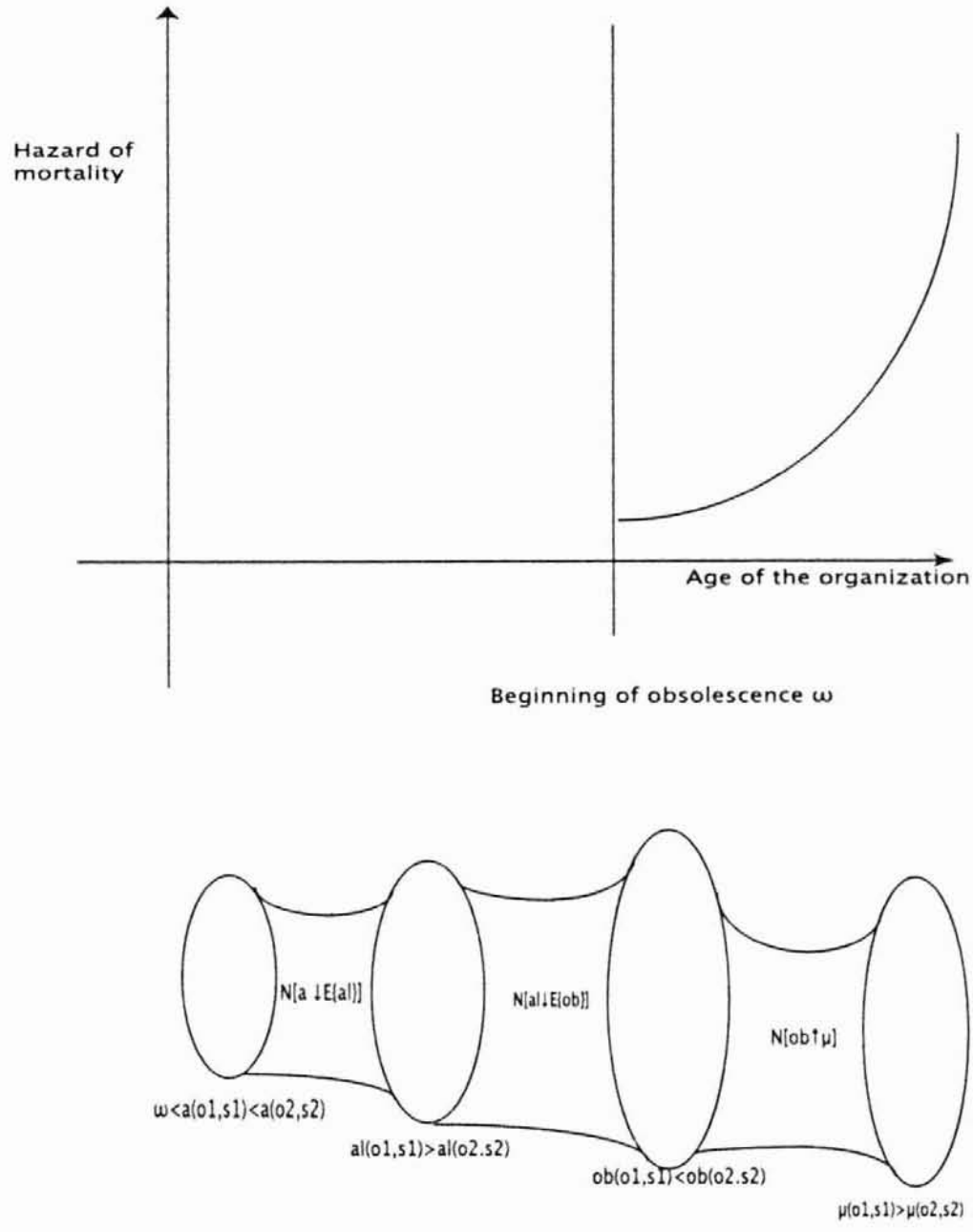

Figure 6. Liability of obsolescence.

rule applies to the end point of an age interval, the hazard increases over the interval.

In this third stage of the theory, the first theorem from the first unification remains valid. Nonetheless, the substantive reasoning behind the theorem has gotten more complex, because we have introduced an obsolescence process. We can illustrate the proof of this theorem in this unified context with a graphical representation of the argument.

Theorem 9: *Mortality hazards increase with age over intervals that begin within expected endowment periods.*

$$\mathfrak{P}o, o', p, s, s'[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \wedge (a_o(s) < \zeta_p) \wedge (a_o(s) < a_{o'}(s')) \rightarrow \mu_o(s) < \mu_{o'}(s')].$$

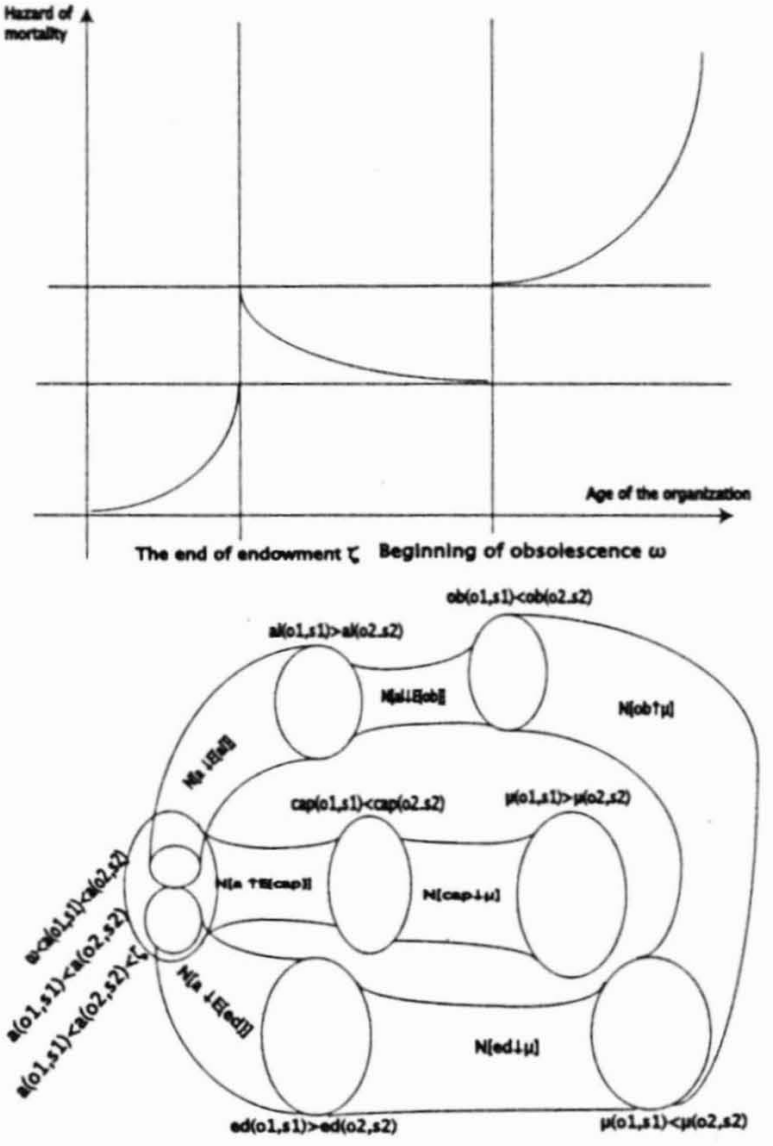

Figure 7. The second unification attempt

Theorem 10: *Mortality hazards decrease over intervals that begin on or after the expected end of the endowment and terminate before the expected*

onset of obsolescence.

$$\mathfrak{P}o, o', p, s, s'[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \wedge (\zeta_p \leq a_o(s) < a_{o'}(s') < \omega_p) \rightarrow \mu_o(s) < \mu_{o'}(s')].$$

Only the liability of newness theory is relevant to the intervals that fit the antecedent in this theorem.

Theorem 11: *Mortality hazards increase over intervals that end at or after the expected onset of obsolescence.*

$$\mathfrak{P}o, o', p, s, s'[\mathrm{O}(o,p) \wedge \mathrm{O}(o',p) \wedge (a_o(s) < \omega_p \leq a_{o'}(s')) \rightarrow \mu_o(s) < \mu_{o'}(s')].$$

Again we can derive implications about jumps and maxima in the process — but only when obsolescence follows the end of endowment.

Theorem 12: *When the expected onset of obsolescence does not occur after the expected end of endowment ($\omega_p \leq \zeta_p$), an organization's mortality hazard presumably jumps at the end of endowment and at the onset of obsolescence.*

Our model of the local behavior of the process yields an unexpected pattern: global positive age dependence. Two cases need to be considered. In the simpler case, when obsolescence strikes before endowments end at the same time ($\zeta_p \leq \omega$), then mortality hazards *increase* with age at all ages.

The second, more complex case, involves a delay between the ending of endowment and the onset of obsolescence ($\zeta_p < \omega_p$). Inspection of Figure 7 reveals an age range in which the default does not get overridden. So there is a period in which the hazard falls with increasing aging. But the hazard over this range must always exceed the maximum hazard during endowment. The overall pattern for this case has the general form shown in Figure 7.

Under specific conditions, the general picture reproduces the patterns of age-dependence found in empirical research, as can be seen by consulting Figure 7.

i. If the organizations in a population lack endowments and occupy environments that change so gradually that obsolescence never strikes, then the default never gets overridden: age dependence is presumably uniformly negative.

ii. If the exhaustion of endowments does not occur within an observation period or obsolescence strikes before exhaustion of endowments, then age dependence is presumably uniformly positive.
iii. If the organizations in a population are endowed and do not face obsolescence, then the mortality hazard presumably peaks in adolescence.
iv. If the organizations in a population are endowed and do face obsolescence at a time later than the ending of endowment, then the mortality hazard presumably has the age profile illustrated in Figure 7

Conclusion: Critical Challenges to Empirical Theories

In this paper we offered a formal semantic account on the argumentation in theory building. In the appendix we briefly summarized some formalizations that are based on this logic, and showed that a nonmonotonic rendering of the argumentation not only allows for otherwise impossible theory unification, but yield relevant substantive insights too. Encouraged by the positive results of these application attempts we try to sketch what does it mean for some problems in the philosophy of science if accepts that our rendering of the argumentation in theory building is correct.

Although rules-with-exceptions might be true or false, their truth and falsity is not expressible in terms of truth and falsity of the corresponding sentences about individual instances. The rules are false if the regularity they express is not present in world. If one can show that indeed this is the case, then the theory is falsified, and it should be discarded as Popper argued. However the falsifiability, according to our rendering still helps to solve the demarcation problem, it appears to be the criterion that discriminates between the substantive core of a theory and the protective belt around it. On the other hand the end of a scientific research program, a theory in our terminology, is not as sharp as Popper envisaged it. Think, for example of the generalization that "In the case of burning, phlogiston leaves the burning material." This statement happens to be false; but the proof of its falsity did (and could) not happen by finding counter examples, cases of burning where phlogiston did not leave the burning material. However, proving the *absence* of a regularity is not any easier than proving its *presence*.

Lakatos argued that the prolonged agony of failing research programs is due to the protective belt which generates problem shifts and this process does not stop even when the problem shifts are frequently negative. We argued that the protective belt is not the only responsible party in this situation.

Predictions are built from generic rules, and they might turn out to be false. Since predictions can be about individual instances, proving the falsity of such predictions might be, perhaps, an easier task. But even when such

a move succeeds, the theory is not discarded, and it should not be. False predictions indicate that there are exceptions to the regularities, but that is not unexpected. Yet, finding the actual exceptions can have an impact on the theory. The discovery of exceptions indicates that the set of explanatory principles the theory provided so far is incomplete, or alternatively further auxiliary assumptions are needed to eliminate the inconsistency. The core of the theory has to be extended with causal considerations that help to construct *more specific* arguments concerning the individual instance in question, or alternatively the protective belt should expand. It is easy to see that neither of these changes does come cheap. Both increases the intricacy of the theory.

Due to the nonmonotonic nature of the argumentation, one has to consider all the assumptions, postulates in every proof, and it becomes increasingly harder to develop a vision of what might be true (what might be provable) in a theory in flux. If the positive problem shifts are the rewards the researchers get for their efforts to cultivate a theory, the intricacy of the theory might be seen as the cost to cultivate. Of course, there is no reason to believe that decisions about what research program should an individual researcher follow are typically rational. Often they are not. However, those who pay high costs for low returns might not become successful researchers, might even leave the field. Theories and research programs might close for the reason that there is not sufficient concentration of brain power left to protect the core successfully. If this description of potential failure of scientific research programs is correct, it explains why theories normally do not vanish before a viable alternative appears on the scene, one that is able to recruit new, and capable defenders. And it might be easier to recruit new defenders if the core of the theory is less intricate.

László Pólos
University of Durham

Michael T. Hannan
Stanford University

REFERENCES

Barwise, Jon and John Perry. 1983. *Situations and Attitudes.* Cambridge: MIT Press.

van Benthem, Johan F.A.K. 1996. "Logic and Argumentation Theory." In *Logic and Argumentation,* edited by J. van Benthem, S. van Eemereren, R. Grootendorst, and F. Veltman. Amsterdam: Royal Dutch Academy of Sciences.

Brewka, G., J. Dix, and K. Konolige. 1997. *Nonmonotonic Reasoning: An Overview.* Stanford, Cal.: CSLI Publications.

Carlsson, Gregory N. 1977. *Reference to Kind in English.* Amherst: Ph.D. Dissertation University of Massachusetts.

Carlsson, Gregory N. 1995. "Truth-Conditions of Generic Sentences: Two Contrasting Views." Pp. 224–37 in G.N. Carlson and F.J. Pelletier (eds.) *The Generic Book.* Chicago: University of Chicago Press.

Diesing, Molly. 1995. "Bare Plural Subjects and the Stage/Individual Contrast." Pp. 107–154 in *The Generic Book.* Chicago: University of Chicago Press.

Duhem, P. 1906 *La théorie physique, son objet et sa structure* English translation: *The Aim and Structure of Physical Theory* Princeton University Press, 1954.

Hannan, Michael T., Glenn R. Carroll, and László Pólos. 2003a. "The Organizational Niche." *Sociological Theory* 21:309–40.

Hannan, Michael T., Glenn R. Carroll, and László Pólos. 2003b. "A Formal Theory of Resource Partitioning." Research paper 1763. Stanford Graduate School of Business.

Hannan, Michael T., László Pólos, and Glenn R. Carroll. 2003a. "Cascading Organizational Change." *Organization Science* 14:463–82.

Hannan, Michael T., László Pólos, and Glenn R. Carroll. 2003b. "The Fog of Change: Opacity and Asperity in Organizations." *Administrative Science Quarterly* 48:399–432.

Hannan, Michael T., László Pólos, and Glenn R. Carroll. 2004. "The Evolution of Inertia." *Industrial and Corporate Change* 13: 213–42.

Hannan, Michael T., László Pólos, and Glenn R. Carroll. 2005. *Social Codes and Ecologies: A Treatise on Organizations.* draft book manuscript.

Kuhn, Thomas S. 1996. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Kratzer Angelika. 1995. "Stage Level and Individual Level Predicates." Pp. 125–175 in G. Carlsson and D.F. Pelletier (eds.) *The Generic Book.* Chicago: University of Chicago Press.

Lakatos, Imre. 1987. *Proofs and Refutations: The Logic of Mathematical Discovery.* Cambridge: Cambridge University Press.

Lakatos, Imre. 1994. *The Methodology of Scientific Research Programmes: Philosophical Papers, Vol. 1.* Cambridge: Cambridge University Press.

Makinson, D. 1994 "General Nonmonotonic Logic." Pp. 35–110 in D.M. Gabbay, C.J. Hogg, and J.A. Robinson (eds.) *Handbook of Logic in Artificial Intelligence and Logic Programming: Nonmonotonic Reasoning and Uncertain Reasoning.* Vol. III. Oxford: Oxford University Press.

Pólos, László, Michael T. Hannan, and Jaap Kamps. 1999. "Aging by Default." Pp. 207–219 in H. Rott, C. Albert, G. Brewka and C. Wittveen

(eds.) *Proceeding of the Fourth Dutch-German Workshop on Non-Monotonic Reasoning Techniques and Their Applications.* Amsterdam: ILLC.

Pólos, László and Michael T. Hannan. 2001. "Nonmonotonicity in Theory Building." Pp. 405–38 in A. Lomi and E. Larson (eds.) *Dynamics of Organizations: Computational Modeling and Organization Theories.* Cambridge: AAI/MIT Press.

Pólos, László and Michael T. Hannan. 2002. "Reasoning with Partial Knowledge." *Sociological Methodology* 32:133–181.

Pólos, László, Michael T. Hannan, and Glenn R. Carroll. 2002. "Foundations of a Theory of Social Forms." *Industrial and Corporate Change* 11:85–115.

Popper, Karl. 1959. *The Logic of Scientific Discovery.* New York: Basic Books.

Popper, Karl. 1963. *Conjectures and Refutations.* London: Routledge & Kegan Paul.

Schubert, Lenhardt, and Francis J. Pelletier. 1988. "An Outlook on Generic Sentences." Pp. 357–372 in M. Krifka (ed.) *Genericity in Natural Language: Proceedings of the 1988 Tübingen Conference.* . Tübingen: Universität Tübingen.

Stinchcombe, Arthur S. 1965. "Social Structure and Organizations." Pp. 142–93 in J.G. March (ed.) *Handbook of Organizations.* Chicago: Rand McNally.

Veltman, Frank. 1991. "Defaults in Update Semantics." Technical Report LP 91-02 ILLC, Universiteit van Amsterdam.

Veltman, Frank. 1996. "Defaults in Update Semantics." *Journal of Philosophical Logic* 25:221–61.